# Performance Analysis of MPI over InfiniBand on Yellowstone

Zhengyang Liu
Mentor: Dr. John Dennis

Collaborators:
Prof. Malathi Veeraraghavan (University of Virginia)
Prof. Robert D. Russell (University of New Hampshire)
Fabrice Mizero (SIParCS)
Patrick MacArthur (University of New Hampshire)

Oct 25, 2013

CISL Computational & Information Systems Laboratory

NCAR UNIVERSITY of VIRGINIA

## Big Picture

- **Understanding the causes of poor performance of CESM on Yellowstone: a 5-step approach**
  - Experimental execution and data collection
  - CESM trace analysis
  - IBMgtSim: routing study
  - Network simulation
  - Integrated simulation

CISL Computational & Information Systems Laboratory

NCAR UNIVERSITY of VIRGINIA

1

# Big Picture

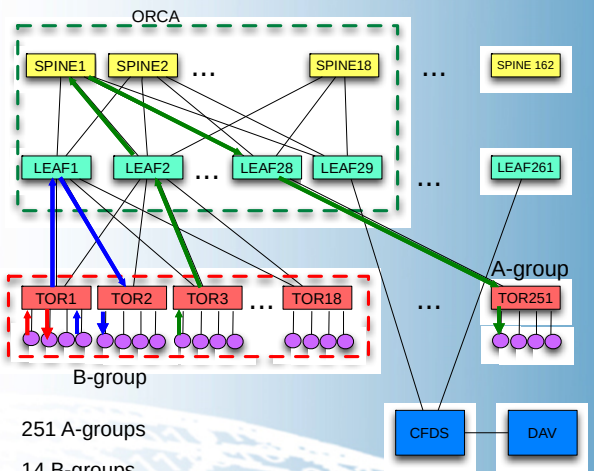
- **Understanding the causes of poor performance of CESM on Yellowstone: a 5-step approach**
  - Experimental execution and data collection
  - CESM trace analysis
  - IBMgtSim: routing study
  - Network simulation
  - Integrated simulation

Computational & Information Systems Laboratory

NCAR UNIVERSITY of VIRGINIA



## Yellowstone network

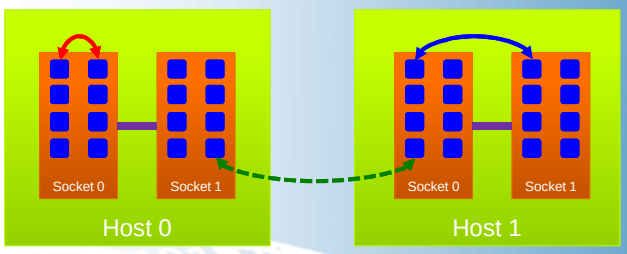



*Credit: Dr. John Dennis

Computational & Information Systems Laboratory

NCAR UNIVERSITY of VIRGINIA

# Communication Patterns

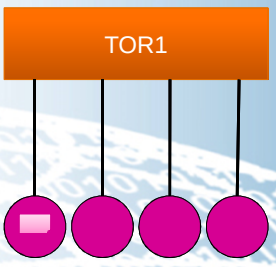



Intra-socket (via shared cache/memory)
Inter-socket (via shared memory over QuickPath Interconnect)
Inter-node (via InfiniBand)

Computational & Information Systems Laboratory
CISL
NCAR
UNIVERSITY of VIRGINIA
4

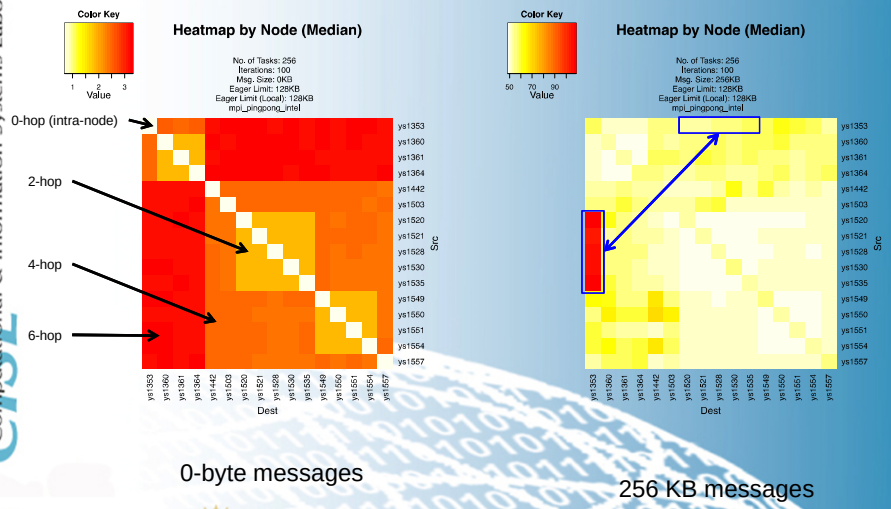
# Latency Benchmark: mpi_pingpong

- Approximate one-way latency by measuring round-trip latency
- Results represent ideal latencies between nodes




Computational & Information Systems Laboratory
CISL
NCAR
UNIVERSITY of VIRGINIA
5

# Yellowstone Results




0-byte messages

256 KB messages



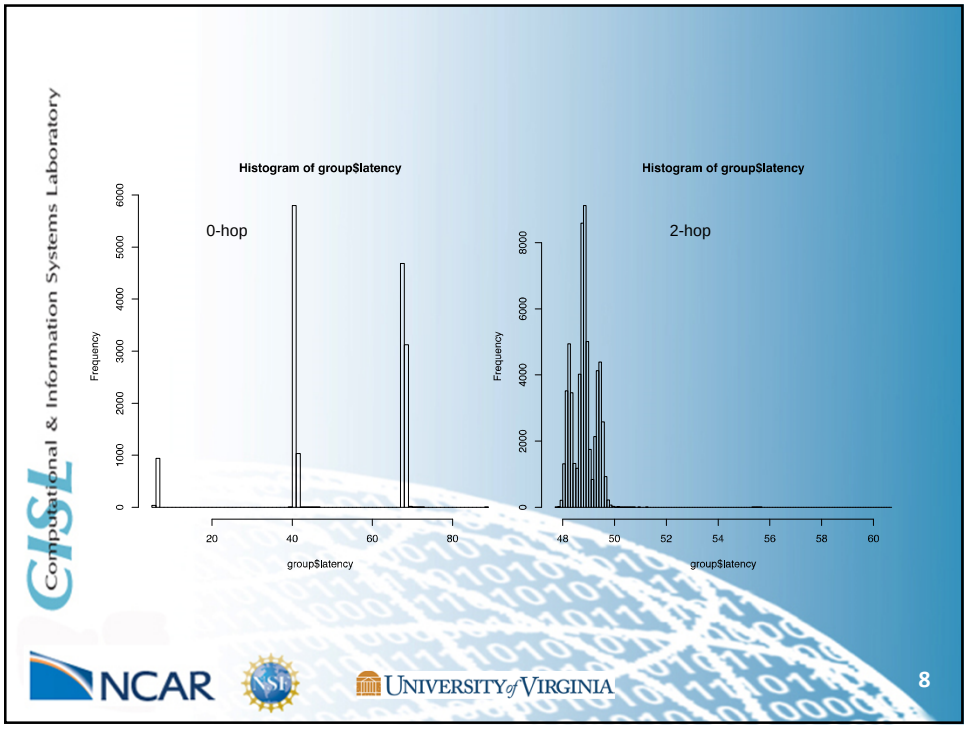
# Latency vs. # of Hops

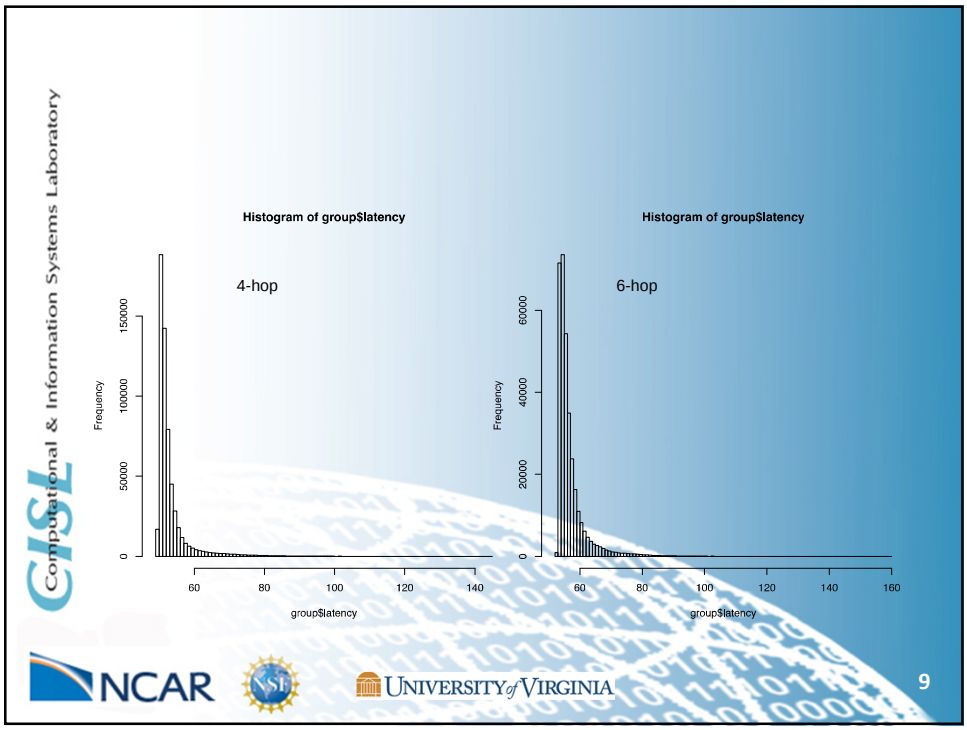
- **Experiment:**
  - mpi_pingpong on 1024 cores
  - 1,048,576 communication pairs*
  - 256 KB messages

Unit: μs

| # of Hops | # of pairs | Min. | Avg. | Max. |
|---|---|---|---|---|
| 0 | 15,680 | 5.95 | 52.20 | 88.29 |
| 2 | 59,904 | 47.72 | 48.83 | 60.64 |
| 4 | 588,736 | 49.55 | 53.30 | 114.10 |
| 6 | 332,016 | 52.75 | 56.92 | 159.30 |

*routing data available for 996,336 pairs


Computational & Information Systems Laboratory
CISL
Histogram of group$latency
0-hop
Frequency
group$latency
Histogram of group$latency
2-hop
Frequency
group$latency
NCAR
UNIVERSITY of VIRGINIA
8


Computational & Information Systems Laboratory
CISL
Histogram of group$latency
4-hop
Frequency
group$latency
Histogram of group$latency
6-hop
Frequency
group$latency
NCAR
UNIVERSITY of VIRGINIA
9

# Bandwidth Benchmark: mpi_bw

- **Measures throughput between two MPI ranks**
- **3 communication patterns:**
  - Intra-socket
  - Inter-socket
  - Inter-node
- **2 communication protocols:**
  - Eager protocol
  - Rendezvous protocol

CISL Computational & Information Systems Laboratory

NCAR NSF UNIVERSITY of VIRGINIA

10

# Communication Protocols

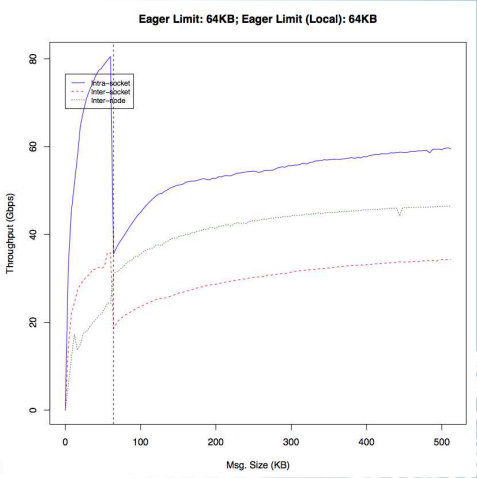
- **Rendezvous Protocol: buffer negotiation before sending**
- **Eager Protocol: send directly without confirming available buffer space**
- **InfiniBand: Eager protocol uses SEND/RECV verbs (two-sided communication); Rendezvous protocol uses WRITE/READ verbs (one-sided communication)**
- **Eager Limit: threshold below which Eager protocol is used**

CISL Computational & Information Systems Laboratory

NCAR NSF UNIVERSITY of VIRGINIA

11

# Jellystone Results




- Intra-node throughput decreases when msg. size > eager limit
- Inter-node throughput increases when msg. size > eager limit
- Inter-node communication faster than inter-socket communication: RDMA vs shared memory

Computational & Information Systems Laboratory

NCAR UNIVERSITY of VIRGINIA

12

# IBMgtSim and OpenSM

- **Problem**
  - Special version of OpenSM needed to boot a simulated network
  - Compile time configuration, not supported by Mellanox's version
  - Open source version of OpenSM does not work with Mellanox IBUtils (which contains IBMgtSim)
  - Open source IBUtils outdated
- **Solution**
  - Fix bugs in open source version of OpenSM
  - Patches OpenSM to support FDR link rate
  - Patches to support better simulation work flow (RunSimTest)

Computational & Information Systems Laboratory

NCAR UNIVERSITY of VIRGINIA

13

# Summary

- **Identified contention through mpi_pingpong benchmarks**
- **Studied effect of different communication patterns/protocols on throughput**

CISL Computational & Information Systems Laboratory

NCAR NSF UNIVERSITY of VIRGINIA

14

# Future Work

- **Analyses of larger data sets**
  - > 500 million data points
  - Analysis needs to be parallelized
- **Study interaction between MPI and InfiniBand**
  - Open-source MPI implementations
  - Network sniffing

CISL Computational & Information Systems Laboratory

NCAR NSF UNIVERSITY of VIRGINIA

15


CISL
Computational & Information Systems Laboratory
Thank You
Zhengyang Liu
zl4ef@virginia.edu
NCAR
UNIVERSITY of VIRGINIA